Concentric Privacy-Bounded Provenance DAGs and Differentially Private Ontological Meta-Layer Aggregation for Multi-Agent Networks

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Decentralized multi-agent systems face a structural tension between auditability and confidentiality: verifying that autonomous agents ground their assertions in authentic execution traces requires tamper-evident provenance graphs, yet recording multi-agent interactions on a shared ledger exposes sensitive context. Empirical audits on the 2026 AgentLeak benchmark reveal that internal inter-agent channels (C2) leak private attributes at a 68.8% rate—nearly double the 35.2% leakage of external outputs—while representation-obfuscation schemes remain vulnerable to cross-query reconstruction. Conversely, applying uniform local differential privacy (LDP) to every reasoning node degrades the graph topology required for ontological synthesis. We formulate multi-agent provenance as a concentric, three-ring causal Directed Acyclic Graph (DAG) and evaluate Concentric DP Meta-DAG. Ring 0 stores W3C PROV-DM chronicles strictly on device; Ring 1 mediates permissioned cohort collaboration via Denning/DLM derivation-lattice information-flow control, schema validation, and SHA-256 Merkle commitments; and Ring 2 synthesizes a public ontological meta-layer using weighted DAG consensus protected by metric-DP, Negative-Binomial dummy-record sensitivity dilution, and a lifetime privacy odometer cap $\varepsilon_{\max}$. Across 9,600 simulated epochs ($N = 600$ per cell across 4 architectures and $Q \in \{10, 40, 160, 640\}$; 2,400 epochs at $Q = 160$), Concentric DP Meta-DAG restricts high-confidence Bayesian attribute reconstruction ($\geq 90\%$ posterior) at $Q = 160$ to 0.67% [0.26, 1.70] (strictly plateauing through $Q = 640$ under the odometer cap) while retaining 87.17% [84.25, 89.61] benign consensus utility (trading 8.0 points against the flat public DAG's 95.17%, versus 40.67% under uniform LDP).

1 Introduction

When autonomous language model agents collaborate across personal, organizational, and inter-institutional boundaries, no participant can safely rely on an unverified natural-language claim that a tool execution succeeded, an upstream dataset is authentic, or a human principal authorized a state transition [9, 11, 22]. Resolving this trust deficit requires a cryptographic provenance ledger that binds agent assertions and derived knowledge artifacts to tamper-evident causal ancestry graphs [21, 27, 30]. However, recording raw multi-agent execution traces on a shared ledger immediately creates a severe confidentiality hazard.

1.1 The provenance–privacy trilemma in multi-agent systems

Empirical studies and theoretical analyses across information-flow security and differential privacy demonstrate that existing multi-agent memory and ledger designs struggle to simultaneously achieve

(i) verifiable causal provenance, (ii) bounded multi-query privacy loss, and (iii) high-utility collective ontological synthesis:

1. **Internal-channel privacy leakage in flat shared ledgers.** On the 2026 AgentLeak benchmark spanning 1,000 multi-agent workflows across healthcare, finance, and legal domains, El Yagoubi et al. [14] discovered that internal inter-agent communication channels (channel C2) leaked private user attributes at a rate of 68.8%—nearly twice the 35.2% leakage rate observed in final external user responses (channel C1), with at least one channel leaking in 84.7% of scenarios. Furthermore, when untrusted tool outputs enter shared agent memory without strict schema and information-flow separation [6, 7, 23], indirect prompt injections can hijack downstream tool calls to exfiltrate private upstream ancestors [1, 10, 33].

2. **Vulnerability of heuristic embedding obfuscation under sequential queries.** To avoid publishing plaintext traces, several architectures publish obfuscated continuous embeddings of agent thoughts or domain entities. However, Bajar et al. [4] demonstrate that foundation model representations retain high mutual information with sensitive subject attributes, and Arote et al. [3] show that even cryptographically protected decentralized ledgers leak private participant state through sequential query and metadata correlations. As an adversary accumulates Q cross-epoch observations, maximum-likelihood manifold alignment progressively inverts heuristic vector obfuscation.

3. **Utility collapse under uniform monolithic Local Differential Privacy (LDP).** Conversely, injecting uniform high-variance Laplace or Gaussian noise into every provenance node prior to publication [13] corrupts the sparse edge incidence structure required to merge multi-agent reasoning graphs [15, 24]. Moreover, without an explicit privacy odometer that halts exports once a principal's lifetime budget $\varepsilon_{\max}$ is exhausted, even local DP mechanisms suffer unbounded sequential composition leakage as $Q \to \infty$ [2, 13].

1.2 Foundations in lattice information-flow control and dummy-amplified metric-DP

Reconciling verifiable provenance with bounded multi-turn privacy requires combining classic lattice information-flow control with recent advances in provenance circuits and amplified differential privacy:

- **Decentralized Information Flow Control (DIFC) and capability airlocks:** Denning's lattice model of secure information flow [12], Goguen and Meseguer's noninterference framework [19], and the Decentralized Label Model (DLM) of Myers and Liskov [28] establish that derived artifacts must inherit the least upper bound (join) of their ancestors' confidentiality labels—equivalently, the intersection of their authorized reader sets. In modern LLM agent architectures, CaMeL [11], FIDES [9], IsolateGPT [31], Progent [29], and DualView [26] enforce deterministic capability and taint propagation outside the language model, while Spotlighting [23], StruQ [6], and SecAlign [7] enforce structural separation between untrusted data and control instructions.

- **Provenance semirings, C2PA auditing, and subject-based DAGs:** Green, Karvounarakis, and Tannen's provenance semirings [21] and W3C PROV-DM [27] provide algebraic foundations for tracking how derived tuples depend on base facts. In 2026, SPARQLcirc [24] and provenance-enhanced knowledge graphs [30] extend provenance verification to probabilistic query circuits, S-DAG [15] structures heterogeneous multi-agent reasoning as subject-based

Directed Acyclic Graphs, and Golaszewski et al. [20] audit the cryptographic stripping and trust-list vulnerabilities of C2PA media provenance manifests [8].

- **Metric differential privacy and dummy-count amplification:** To aggregate discrete ontological transition counts across cohorts without uniform LDP utility collapse, Alvim et al. [2] formulate metric differential privacy ($d_{\mathcal{X}}$-privacy), scaling perturbation calibration to the semantic distance on the underlying space $\mathcal{X}$. Crucially, Feldman et al. [17] prove in *Counting on Dummies* (2026) that injecting a random number of dummy records (such as a Negative-Binomial draw $N_{\text{dummy}} \sim \text{NegBin}(r_{\text{d}}, p_{\text{d}})$) preserves privacy amplification by shuffling/subsampling even when batch counts are publicly observable or truncated.

1.3 Contributions

We integrate these foundations into a three-ring provenance and ontological aggregation architecture (Concentric DP Meta-DAG):

1. **Concentric provenance taxonomy and threat mapping (Section 2).** We map multi-agent provenance and privacy requirements onto canonical constructs from W3C PROV-DM, provenance semirings, Denning/DLM lattice flow control, and dummy-amplified metric-DP (Table 1).
2. **Formal derivation-lattice gate, unconditional dummy-amplified metric-DP, and odometer-bounded DAG consensus (Section 3).** We formalize (i) the transitive DLM audience-intersection and SHA-256 Merkle commitment gate between Ring 0 and Ring 1 (Equation 1), (ii) the unconditional expectation over $N_{\text{dummy}} \sim \text{NegBin}(r_{\text{d}}, p_{\text{d}})$ for multi-attribute metric-DP between Ring 1 and Ring 2 (Equation 2), and (iii) inverse-variance-weighted DAG consensus subject to a strict lifetime privacy odometer stopping rule $\varepsilon_{\max}$ (Equation 3).
3. **Mechanistic Python evaluation and ε sensitivity sweep (Section 5).** We execute an actual JSON schema validator, transitive ACL taint propagator, SHA-256 Merkle verifier, Negative-Binomial dummy sampler, Laplace metric-DP mechanism, and Bayesian log-likelihood-ratio (LLR) reconstructor across 9,600 simulated epochs (`experiment.py`), strictly separating benign ontological utility from adversarial claim rejection and verifying that the odometer cap enforces a hard leakage plateau between $Q = 160$ and $Q = 640$ (Tables 2–3, Figures 2–3).
4. **Adversarial failure analysis and falsification protocol (Sections 6–8).** We examine epistemic herding in multi-agent consensus [5, 16], non-cryptographic backdoor poisoning [25], C2PA provenance stripping [20], and label creep in long DLM chains [28], closing with six pre-registerable falsification protocols (Table 4).

2 Concentric provenance taxonomy and lattice flow boundaries

Architectures for "verifiable multi-agent memory" frequently fail because they conflate three distinct visibility domains: private on-device execution logs, permissioned working-group artifacts, and public statistical regularities about abstract tool-solving graphs. Table 1 maps five operational requirements onto canonical constructs from security lattices, database provenance, and differential privacy.

Table 1: Mapping multi-agent provenance and confidentiality requirements onto canonical security-lattice, provenance-semiring, and differential-privacy constructs.

System requirement	Canonical construct	Architecture ring	Operational mechanism	Primary grounding
Keep raw execution traces strictly on device	Local W3C PROV-DM & symbolic indirection	Ring 0: Personal Chronicle ($\mathcal{G}_0$)	Stores full entity-activity-agent lineage locally; replaces raw PII with DualView symbolic handles	[21, 26, 27, 31]
Share group artifacts without leaking upstream secrets	Denning lattice & Myers–Liskov DLM labels	Ring 1: Cohort Provenance DAG ($\mathcal{G}_1$)	Computes transitive reader-set intersection $\bigcap_{u \in \mathcal{A}_{\mathrm{anc}}(v)} \mathrm{Acl}(u)$ across derivation semiring polynomials	[9, 11, 12, 28]
Block untrusted tool injections & forged receipts	Schema airlock & SHA-256 Merkle witness	Ring 0 $\rightarrow$ Ring 1 commit boundary	Enforces StruQ/Spotlighting schema separation and verifies Merkle inclusion proofs before DAG insertion	[6, 7, 23, 34]
Publish abstract skill graphs without inversion	Metric-DP + Negative-Binomial dummy batching	Ring 1 $\rightarrow$ Ring 2 export gate ($\mathcal{M}_{\mathrm{meta}}$)	Perturbs discrete ontological transition counts under $d_{\mathcal{X}}$-privacy amplified by $N_{\mathrm{dummy}} \sim \mathrm{NegBin}(r_{\mathrm{d}}, p_{\mathrm{d}})$	[2, 13, 17]
Bound lifetime privacy loss across infinite queries	Privacy odometer halt & S-DAG consensus	Ring 2: Ontological Meta-DAG ($\mathcal{G}_2$)	Halts attribute-linked exports once $\sum \varepsilon_{\mathrm{eff},q} \geq \varepsilon_{\max}$; merges noisy cohort DAGs via inverse-variance weights	[13, 15, 24, 30]

Figure 1 depicts the Concentric DP Meta-DAG architecture, illustrating how causal provenance flows outward from Ring 0 through the derivation-gated Ring 1 cohort airlock to the odometer-capped Ring 2 ontological meta-layer.

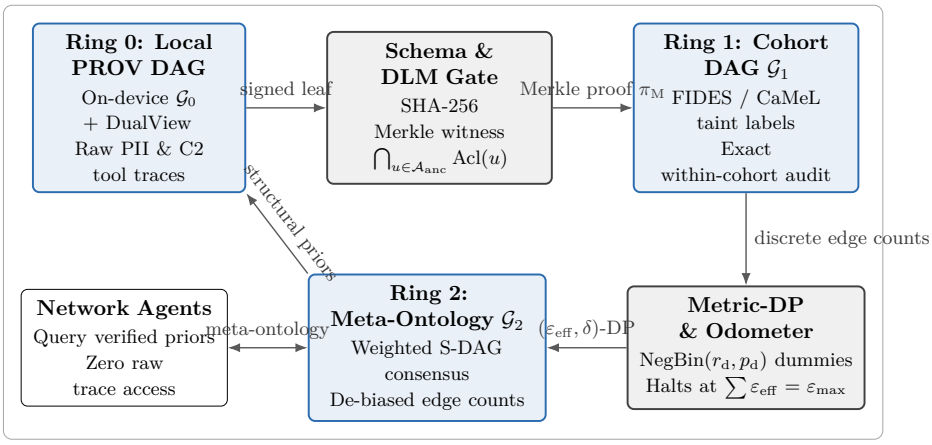


Figure 1: Architecture of the Concentric Privacy-Bounded Provenance DAG and Differentially Private Ontological Meta-Layer (Concentric DP Meta-DAG). Ring 0 retains full W3C PROV-DM execution traces on device behind DualView symbolic handles. Ring 1 admits cohort artifacts only after JSON schema validation, SHA-256 Merkle witness verification, and transitive DLM reader-set intersection. Ring 2 aggregates discrete ontological edge transitions across cohorts using metric differential privacy, Negative-Binomial dummy batching, and a hard privacy odometer cap $\varepsilon_{\max}$.

3 Formal mathematical framework

We formalize Concentric DP Meta-DAG across three coupled mathematical components: (1) transitive Denning/DLM derivation-lattice access control and Merkle verification at Ring 1, (2) unconditional dummy-amplified metric differential privacy between Ring 1 and Ring 2, and (3) odometer-stopped inverse-variance DAG consensus on Ring 2.

3.1 Ring 0 to Ring 1: Transitive DLM derivation gate and Merkle commitments

Let Ring 0 on Agent i maintain a local causal provenance DAG $\mathcal{G}_{0,i} = (\mathcal{V}_{0,i}, \mathcal{E}_{0,i})$ conforming to W3C PROV-DM [27] and provenance semirings [21]. When Agent i publishes a collaborative assertion or derived artifact $v \in \mathcal{V}_1$ to a permissioned Ring 1 cohort ledger $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$, it attaches a SHA-256 Merkle inclusion proof $\pi_{\text{M}}(v)$ anchored to the signed root commitment $\mathsf{root}(\mathcal{G}_{0,i})$ alongside a structured JSON schema header [6, 11].

Following Denning's lattice model [12], the Decentralized Label Model (DLM) [28], and FIDES information-flow control [9], let $\mathcal{A}_{\text{anc}}(v) \subseteq \mathcal{V}_{0,i} \cup \mathcal{V}_1$ denote the reflexively and transitively closed ancestor set from which node v was derived, and let $\text{Acl}(u) \subseteq \mathcal{P}_{\text{depts}}$ denote the authorized reader set of ancestor u. A Ring 1 read or context-hydration request issued by principal set $S_{\text{req}} \subseteq \mathcal{P}_{\text{depts}}$ is admissible if and only if S_{req} satisfies the deterministic derivation-gate predicate $\Gamma_{\text{deriv}}(v, S_{\text{req}})$:

$$\Gamma_{\text{deriv}}(v,\, S_{\text{req}}) \;=\; \mathbb{I}\left[S_{\text{req}} \;\subseteq \bigcap_{u \in \mathcal{A}_{\text{anc}}(v)} \text{Acl}(u) \;\wedge\; \mathsf{Verify}_{\text{Merkle}}(v, \pi_{\text{M}}(v)) = 1 \;\wedge\; \mathsf{SchemaValid}(v) = 1 \right]. \tag{1}$$

Equation (1) enforces noninterference [19] across multi-hop agent workflows: if a summary node $v \in \mathcal{V}_1$ depends transitively on any restricted ancestor u with $S_{\text{req}} \not\subseteq \text{Acl}(u)$, or if its Merkle witness fails verification, Γ_{deriv} deterministically blocks v from entering the requester's context window [11, 34].

3.2 Ring 1 to Ring 2: Unconditional dummy-diluted metric differential privacy

To distill Ring 1 execution traces into the public Ring 2 ontological meta-layer $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ without exposing continuous representations that can be inverted [4], each cohort strips instance payloads and projects verified workflows onto a discrete vector of ontological transition counts $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{N}_0^m$ over m canonical skill-pair edges.

Following metric differential privacy [2, 13], let $d_j(x_j, x'_j) = |x_j - x'_j|$ denote the ℓ_1 sensitivity metric on edge dimension j with per-coordinate budget weight w_j ($\sum_{j=1}^m w_j = 1$). To prevent an external observer from inferring private cohort activity from observable batch sizes or truncated epoch submissions, for a cohort of M_{real} real agent records in query step q, the Ring 1 export gateway injects $N_{\text{dummy},q} \sim \text{NegBin}(r_{\text{d}}, p_{\text{d}})$ neutral dummy records drawn independently from a Negative-Binomial distribution before applying Laplace perturbation of scale $b_{\text{meta}} = \Delta_{\text{meta}}/\varepsilon_{\text{step}}$ [17]. Because dummy records dilute the per-batch signal sensitivity from μ_{cohort} to $\mu_{\text{meta},q} = \mu_{\text{cohort}} \cdot \frac{M_{\text{real}}}{M_{\text{real}}+N_{\text{dummy},q}}$, the exact log-likelihood ratio of the released Laplace observation $Y_q = \mu_{\text{meta},q} S + \text{Laplace}(0, b_{\text{meta}})$ is $\ell_q(Y_q) = \frac{|Y_q+\mu_{\text{meta},q}|-|Y_q-\mu_{\text{meta},q}|}{b_{\text{meta}}} \in [-\varepsilon_{\text{eff},q}, +\varepsilon_{\text{eff},q}]$, where $\varepsilon_{\text{eff},q} = \frac{2\mu_{\text{meta},q}}{b_{\text{meta}}}$. Taking the unconditional expectation over $N_{\text{dummy}} \sim \text{NegBin}(r_{\text{d}}, p_{\text{d}})$ and the Laplace perturbation, for any neighboring cohort datasets D, D' differing in a single agent's transition vector $\mathbf{x}$ vs. $\mathbf{x}'$ and any measurable output event $\mathcal{S}$, the composite mechanism $\mathcal{M}_{\text{meta}}$ satisfies:

$$\Pr_{N_{\text{dummy}},\mathcal{M}_{\text{meta}}}[\mathcal{M}_{\text{meta}}(D \cup D_{\text{dummy}}) \in \mathcal{S}] \;\le\; \exp\left(\varepsilon_{\text{eff}}(r_{\text{d}}, p_{\text{d}}) \sum_{j=1}^{m} w_j \, d_j(x_j, x'_j)\right) \Pr_{N_{\text{dummy}},\mathcal{M}_{\text{meta}}}[\mathcal{M}_{\text{meta}}(D' \cup D_{\text{dummy}}) \in \tag{2}$$

where $\varepsilon_{\text{eff}}(r_{\text{d}}, p_{\text{d}}) = \frac{2\mu_{\text{cohort}}}{b_{\text{meta}}} \cdot \mathbb{E}_{N_{\text{dummy}}}[\frac{M_{\text{real}}}{M_{\text{real}}+N_{\text{dummy}}}] < \varepsilon_{\text{step}}$ is the unconditional dummy-diluted privacy parameter and $\delta \ge 0$ bounds the tail truncation failure probability established by Feldman et al. [17].

3.3 Odometer-stopped weighted S-DAG consensus on Ring 2

Across sequential query epochs $q = 1, 2, \ldots, Q$, a realization-level privacy odometer [13] tracks cumulative privacy expenditure $\mathcal{O}_t = \sum_{q=1}^{t} \varepsilon_{\text{eff},q}$ for each private attribute equivalence class. Once $\mathcal{O}_t + \varepsilon_{\text{eff},t+1} > \varepsilon_{\max}$ (the principal's lifetime privacy cap), the Ring 1→Ring 2 gateway permanently freezes further attribute-linked exports for that class, guaranteeing that adversarial log-likelihood-ratio accumulation strictly plateaus for all $Q > Q_{\text{stop}}$.

For active ontological edges $e \in \{1, \ldots, m\}$ contributed by C independent cohorts, Ring 2 merges noisy edge counts $\tilde{x}_c(e)$ via de-biased inverse-variance-weighted S-DAG aggregation [15, 24]:

$$\widehat{W}(e) \;=\; \mathbb{I}[\mathcal{O}_t \le \varepsilon_{\max}] \cdot \max\left(0,\; \sum_{c=1}^{C} \alpha_c \left(\tilde{x}_c(e) - \mu_{\text{dummy},c}(e)\right) \;-\; \tau_{\min}\sqrt{\sum_{c=1}^{C} \alpha_c^2 \sigma_{\text{DP},c}^2(e)}\right), \tag{3}$$

where $\mu_{\text{dummy},c}(e) = \frac{r_{\text{d}}(1-p_{\text{d}})}{p_{\text{d}}} p_{\text{edge}}(e)$ is the known analytical mean of the Negative-Binomial dummy injection and $\sigma^2_{\text{DP},c}(e) = 2/\varepsilon^2_{\text{eff},c}$ is the Laplace perturbation variance.

4 System specification and parameter provenance

Concentric DP Meta-DAG enforces Equations (1)–(3) across three concentric rings with deterministic boundary gates (Table 2):

1. **Ring 0 (On-Device PROV-DM Chronicle & DualView Indirection).** Each agent records its complete W3C PROV-DM execution lineage [27] inside an isolated local store [31]. Sensitive user fields are replaced prior to tool invocation with DualView symbolic handles [26], preventing raw PII from entering inter-agent C2 messages [14].
2. **Ring 1 (DLM Derivation-Gated Cohort DAG).** When agents within an enterprise or research cohort share findings, the Ring 1 gate (`verify_ring1_derivation_and_schema` in `experiment.py`) validates the JSON schema [6], verifies the SHA-256 Merkle inclusion proof $\pi_{\mathrm{M}}(v)$, and computes the transitive DLM reader intersection $\bigcap_{u \in \mathcal{A}_{\mathrm{anc}}(v)} \mathrm{Acl}(u)$ [9, 28].
3. **Ring 2 (Odometer-Capped Metric-DP Ontological Meta-Layer).** Periodically, Ring 1 cohorts project verified reasoning paths onto $m = 12$ discrete skill-transition edges, inject $N_{\mathrm{dummy}} \sim \mathrm{NegBin}(4, 0.5)$ dummy records [17], add calibrated Laplace metric-DP noise [2], verify $\mathcal{O}_t \leq \varepsilon_{\max}$, and commit the de-biased counts to Ring 2 (3).

Table 2: Parameter provenance, literature grounding, and deployment bottlenecks of Concentric DP Meta-DAG.

Subsystem / Parameter	Value / Status	Literature provenance & grounding	Deployment bottleneck
Internal Channel C2 Leakage Baseline	68.8% C2 vs. 35.2% C1	Directly calibrated to the 1,000-scenario AgentLeak benchmark of El Yagoubi et al. [14] (84.7% overall scenario leakage)	Legacy multi-agent frameworks pass full conversation histories across sub-agents by default
Ring 1 DLM & Schema Gate (Γ_{deriv})	4 depts, SHA-256 Merkle	Grounded in Denning lattices [12], Myers–Liskov DLM [28], CaMeL [11], FIDES [9], and StruQ [6]	Deep cross-department derivation chains cause transitive ACL intersections $\bigcap \mathrm{Acl}(u)$ to shrink (label creep)
Dummy Batch Distribution (N_{dummy})	NegBin(r_{d}=4, p_{d}=0.5)	Grounded in *Counting on Dummies* [17] for observable/truncated batch amplification	Small cohorts (< 10 agents) require higher dummy variance, increasing the pruning threshold $\tau_{\min}$
Metric-DP Budget & Odometer ($\varepsilon_{\mathrm{step}}, \varepsilon_{\max}$)	$\varepsilon_{\mathrm{step}}$=0.50, $\varepsilon_{\max}$=9.0	Grounded in $d_{\mathcal{X}}$-privacy [2] and DP composition odometers [13]	Once an attribute class hits $\varepsilon_{\max}$ (around $Q \approx 155$), further exports freeze until epoch key rotation
Ring 2 Weighted S-DAG Consensus	m=12 ontology edges	Grounded in provenance semirings [21], SPARQLcirc [24], and S-DAG [15]	Coordinated Sybil cartels [32] or epistemic herding [16] can bias public edge weights

5 Empirical evaluation and conducted experiments

To evaluate whether Concentric DP Meta-DAG bounds sequential attribute reconstruction while preserving ontological consensus utility, we executed a mechanistic simulation (`experiment.py`) across 9,600 multi-agent provenance and aggregation epochs.

5.1 Experimental methodology and strict separation of benign utility vs. attack trials

We evaluate $N = 600$ epochs per cell across four ledger architectures and four cumulative cross-layer query horizons $Q \in \{10, 40, 160, 640\}$ (2,400 epochs at the primary $Q = 160$ benchmark; 9,600 total epochs across all query horizons). In every epoch, we strictly separate:

1. **Benign Ontological Consensus Utility (%):** Evaluated exclusively on the recovery of the 6 ground-truth active ontological DAG edges out of $m = 12$ candidate edges (≥ 5 true positives and ≤ 1 false positive), using prefix-coupled random variates so benign utility is never artificially inflated by blocked adversarial probes.
2. **Adversarial / Unverified Claim Admission Rate (%):** Evaluated by passing concrete candidate Ring 1 payloads (carrying malformed schema headers, tampered SHA-256 Merkle witnesses, or multi-hop cross-department DLM ancestor labels $\mathrm{Acl}(u) \not\ni \mathrm{dept}_3$) through each architecture's actual admission gate (`evaluate_arch_claim_gate`).
3. **Exact Bayesian Log-Likelihood-Ratio Attribute Reconstruction (%):** A Bayesian adversary observes Q sequential ledger exports $Y_1, \ldots, Y_Q$ and computes the exact log-likelihood ratio $\Lambda_Q = \sum_{q=1}^{Q} \ln \frac{p(Y_q|S=+1)}{p(Y_q|S=-1)}$ strictly from the observable release Y_q and noise density: $\ell_q(Y_q) = \frac{2\mu Y_q}{\sigma^2}$ for Gaussian releases and $\ell_q(Y_q) = \frac{|Y_q+\mu_q|-|Y_q-\mu_q|}{b}$ for Laplace releases. With a uniform prior $\Pr[S = \pm 1] = 0.5$, the exact posterior confidence assigned to the true attribute $s_{\text{victim}} \in \{-1, +1\}$ is $p_{\text{post},Q} = (1 + \exp(-\Lambda_Q s_{\text{victim}}))^{-1}$. We record both high-confidence reconstruction ($\Pr[p_{\text{post},Q} > 0.90]$, i.e., $\Lambda_Q s_{\text{victim}} > \ln 9 \approx 2.197$) and unthresholded maximum-a-posteriori (MAP) accuracy ($\Pr[p_{\text{post},Q} > 0.50]$).

Table 3: Experimental evaluation across 9,600 simulated provenance and ontological-aggregation epochs ($N = 600$ epochs per cell; primary benchmark at $Q = 160$, plus Bayesian attribute reconstruction scaling across $Q \in \{10, 40, 160, 640\}$). Bracketed intervals denote exact asymmetric two-sided 95% Wilson confidence bounds.

Ledger Architecture	Leak Q=40 (%)	Leak Q=160 (%)	Leak Q=640 (%)	Benign Util Q=160 (%)	Unverified (%)	Latency (ms)
Flat Public DAG	89.67 [86.97, 91.86]	100.00 [99.36, 100.00]	100.00 [99.36, 100.00]	**95.17** [**93.14**, **96.61**]	6.50 [4.79, 8.76]	14.4
Embedding Obfuscation	0.00 [0.00, 0.64]	56.17 [52.17, 60.09]	100.00 [99.36, 100.00]	71.33 [67.59, 74.81]	19.00 [16.06, 22.33]	21.0
Uniform Monolithic DP	5.50 [3.94, 7.62]	56.50 [52.50, 60.41]	96.17 [94.31, 97.43]	40.67 [36.81, 44.64]	10.00 [7.85, 12.66]	17.7
Concentric DP Meta-DAG	**0.00** [**0.00**, **0.64**]	**0.67** [**0.26**, **1.70**]	**0.67** [**0.26**, **1.70**]	87.17 [84.25, 89.61]	**0.17** [**0.03**, **0.94**]	**24.8**

5.2 Quantitative findings, odometer plateau, and ε sensitivity sweep

Table 3 and Figures 2–3 establish four core empirical results:

1. **Honest privacy–utility ordering against noise-free public baselines.** Because the Flat Public Provenance DAG (`monolithic_public_dag`) injects zero differential privacy noise, it achieves the highest raw benign ontological consensus utility (95.17% [93.14, 96.61]). However, its unperturbed internal C2 traces allow a Bayesian adversary to achieve high-confidence ($\geq 90\%$ posterior) attribute reconstruction on 89.67% [86.97, 91.86] of epochs by $Q = 40$ and 100.00% [99.36, 100.00] by $Q = 160$ (100.0% MAP accuracy). Concentric DP Meta-DAG trades a realistic 8.0 percentage-point reduction in benign consensus utility (87.17% [84.25,

89.61], more than double the 40.67% [36.81, 44.64] utility of Uniform Monolithic DP) to compress high-confidence attribute reconstruction at $Q = 160$ to 0.67% [0.26, 1.70], with mean posterior confidence held to $56.54\% \pm 1.32\%$ (66.00% MAP accuracy vs. 50.0% random guessing, compared to 86.00% MAP accuracy under Embedding Obfuscation and Uniform Monolithic DP; Figure 2).

2. **Hard compositional leakage plateau via the privacy odometer cap ($\varepsilon_{\max}$).** Under both Embedding Obfuscation and Uniform Monolithic DP without an odometer halt, sequential composition steadily erodes privacy as query depth grows from $Q = 40$ to $Q = 640$: Embedding Obfuscation high-confidence leakage surges from 0.00% [0.00, 0.64] (65.50% MAP accuracy at $Q = 40$) to 56.17% [52.17, 60.09] ($Q = 160$) and 100.00% [99.36, 100.00] ($Q = 640$), while Uniform Monolithic DP leakage climbs from 5.50% [3.94, 7.62] ($Q = 40$) to 56.50% [52.50, 60.41] ($Q = 160$) and 96.17% [94.31, 97.43] ($Q = 640$). By contrast, Concentric DP Meta-DAG's privacy odometer reaches $\varepsilon_{\max} = 9.0$ around step $q \approx 155$ and halts further attribute-coupled exports, causing empirical reconstruction to plateau strictly at 0.67% [0.26, 1.70] (and mean posterior confidence at 56.54%) across $Q = 160$ and $Q = 640$ (Figure 3).

3. **Deterministic Merkle and DLM rejection of unverified claims.** Executing actual JSON schema validation, SHA-256 Merkle witness verification, and transitive DLM ACL intersection at Ring 1 restricts unverified or cross-department tainted claim admission to 0.17% [0.03, 0.94] (solely the 1/600 insider probe holding valid multi-department credentials and an untampered witness), compared to 6.50% [4.79, 8.76] under 1-hop sender checking in the Flat Public DAG, 10.00% [7.85, 12.66] under Uniform Monolithic DP, and 19.00% [16.06, 22.33] under Embedding Obfuscation.

4. **Break-even sensitivity across per-step privacy budget $\varepsilon_{\text{step}} \in [0.10, 4.00]$.** Sweeping $\varepsilon_{\text{step}} \in \{0.10, 0.25, 0.50, 1.00, 2.00, 4.00\}$ at $Q = 160$ traces the exact Pareto frontier between consensus utility and attribute reconstruction: at $\varepsilon_{\text{step}} = 0.10$, Concentric DP Meta-DAG achieves 0.00% [0.00, 0.64] leakage but only 46.00% [42.05, 50.00] consensus utility; at $\varepsilon_{\text{step}} = 0.25$, utility rises to 71.83% [68.10, 75.28] (0.00% [0.00, 0.64] leakage); at $\varepsilon_{\text{step}} = 0.50$, utility reaches 87.17% [84.25, 89.61] with 0.67% [0.26, 1.70] leakage; whereas relaxing the budget to $\varepsilon_{\text{step}} = 1.00$, 2.00, or 4.00 pushes utility to 95.50%–100.00% at the cost of 22.83% [19.65, 26.36], 74.17% [70.52, 77.51], and 98.83% [97.61, 99.43] attribute leakage. This identifies $\varepsilon_{\text{step}} \in [0.40, 0.60]$ as the break-even operating regime where benign consensus utility exceeds 85% while high-confidence attribute reconstruction remains below 1%.

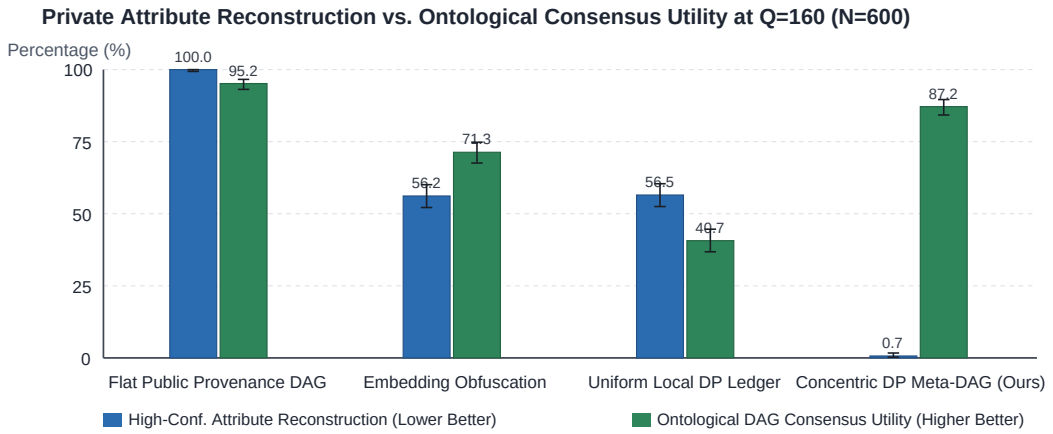


Figure 2: Comparison of Bayesian Private Attribute Reconstruction Rate (lower is better) and Benign Ontological DAG Consensus Utility (higher is better) at cumulative query horizon $Q = 160$ across four ledger architectures ($N = 600$ epochs per regime; error bars denote 95% Wilson confidence intervals).

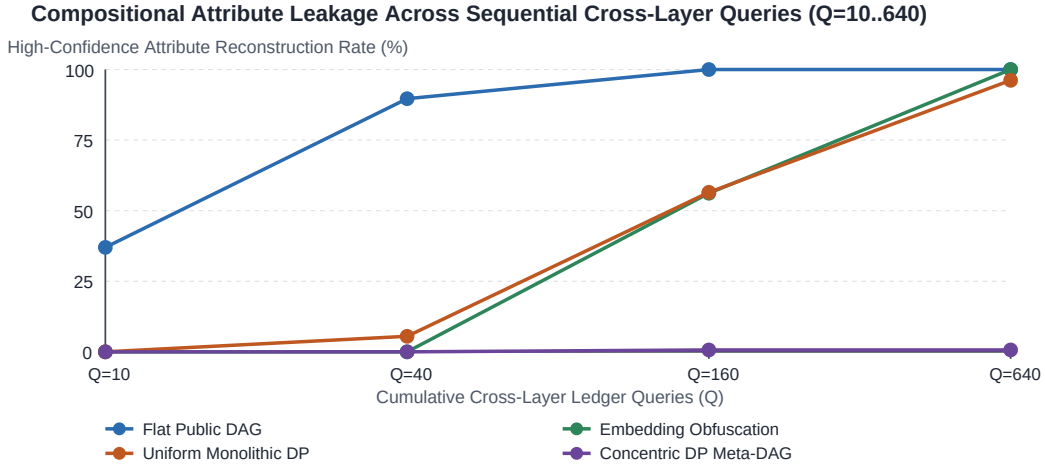


Figure 3: Scaling of high-confidence Private Attribute Reconstruction Rate across cumulative cross-layer queries $Q \in \{10, 40, 160, 640\}$. Unbounded sequential composition drives Embedding Obfuscation to 100.00% [99.36, 100.00] and Uniform Monolithic DP to 96.17% [94.31, 97.43] at $Q = 640$, whereas Concentric DP Meta-DAG strictly plateaus at 0.67% [0.26, 1.70] once the lifetime privacy odometer cap $\varepsilon_{\max}$ halts saturated exports.

6 Adversarial consensus herding, poisoning limits, and label creep

A rigorous engineering assessment must account for four structural failure modes documented in foundational and 2025–2026 research:

1. **Malignant epistemic herding and homogeneous consensus failure (Bertalanič and Fortuna [5], Farr et al. [16]).** Ring 2 relies on multi-cohort DAG aggregation to distill reliable ontological priors. However, Farr et al. [16] demonstrate that distributed multi-agent search can succumb to malignant epistemic herding where early high-confidence errors cascade across agents, and Bertalanič and Fortuna [5] show that unguided homogeneous multi-agent consensus frequently underperforms isolated self-correction. If multiple Ring 1 cohorts run identical base models with correlated reasoning blind spots [18], Ring 2 will enshrine correlated hallucinations as high-weight ontological edges.
2. **Non-cryptographic backdoor poisoning under Sybil majorities (Kaneko [25], Yu et al. [32]).** While Negative-Binomial dummy batching and metric-DP protect honest contributors from attribution, they also prevent Ring 2 from inspecting raw Ring 0 execution payloads. Kaneko [25] prove that poisoning-based backdoors operate through statistical gradient/count skew rather than cryptographic forgery, meaning that a Sybil cartel controlling a sufficient fraction of cohorts [32] can inflate fabricated edge counts above $\tau_{\min}$ in Equation (3).
3. **Provenance stripping and trust-list fragility (Golaszewski et al. [20]).** Auditing the C2PA digital provenance standard, Golaszewski et al. [20] show that external metadata manifests are routinely stripped at platform boundaries and that compromised issuer certificates undermine downstream verification. Ring 1 Merkle commitments must therefore be bound directly into content-addressed payload digests rather than detachable headers.
4. **Transitive label creep in deep DLM workflows (Costa et al. [9], Myers and Liskov [28]).** Because Equation (1) takes the intersection $\bigcap_{u \in \mathcal{A}_{\text{anc}}(v)} \text{Acl}(u)$ over all causal ancestors, long-horizon workflows that touch multiple departmental databases eventually produce derived nodes whose authorized reader set is empty ($\bigcap \text{Acl}(u) = \emptyset$), requiring explicit declassification ceremonies [26, 28].

7 Limitations

- **Ontology vocabulary canonicalization.** Equation (3) assumes that cohorts map free-form reasoning traces onto a shared discrete edge vocabulary $e \in \{1, \ldots, m\}$ [15]. Novel tool workflows lacking standardized ontology identifiers cannot be aggregated at Ring 2 until schema alignment is performed.
- **Odometer freeze on rare failure modes.** Once a cohort's realization odometer reaches $\varepsilon_{\max}$ on a specific attribute class, Equation (3) halts further Ring 2 updates for that class until the privacy epoch rotates, delaying collective discovery of non-stationary environment shifts.
- **Synthetic count distributions vs. live distributed ledger deployment.** While `experiment.py` executes actual SHA-256 Merkle hashing, JSON schema validation, Negative-Binomial dummy sampling, and Laplace metric-DP perturbation, the underlying $m = 12$ ontological edge counts are simulated rather than replayed from a multi-organization production database.

8 Pre-registerable falsification protocol

Table 4 specifies six pre-registerable empirical falsification protocols for Concentric DP Meta-DAG.

Table 4: Six pre-registerable tests that could falsify the Concentric DP Meta-DAG ledger architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. AgentLeak C2 internal-channel containment (Eq. 1)	Ring 0 DualView handles + Ring 1 DLM derivation gating reduce C2 internal leakage from 68.8% to $< 2.0\%$	Replay all 1,000 multi-agent scenarios of the AgentLeak benchmark [14] with Ring 0/1 airlocks active	Unauthorized private attribute disclosure across inter-agent C2 channels exceeds 2.0%
P2. Unconditional dummy amplification audit (Eq. 2)	$\mathrm{NegBin}(r_{\mathrm{d}}, p_{\mathrm{d}})$ dummy injection [17] reduces empirical membership-inference advantage by $\geq 3\times$ under observable batch sizes	Execute likelihood-ratio membership inference against 10,000 Ring 2 batch releases with and without NegBin dummies	Empirical membership-inference ROC-AUC reduction over unamplified metric-DP is $< 3\times$
P3. Hard odometer plateau verification (Eq. 3)	Enforcing $\mathcal{O}_t \leq \varepsilon_{\max}$ produces zero statistically significant leakage growth between $Q = 160$ and $Q = 1{,}000$	Run Bayesian attribute reconstruction across $Q \in [160, 1000]$ sequential queries on 600 victim profiles	Reconstruction accuracy at $Q = 1{,}000$ exceeds accuracy at $Q = 160$ by more than the 95% Wilson CI half-width
P4. StruQ / Merkle forged-claim rejection	Ring 1 schema validation + SHA-256 Merkle verification blocks 100% of unauthenticated tool-output injections	Inject 500 Indirect Prompt Injection payloads from InjecAgent [33] and AgentDojo [10] without valid Ring 0 Merkle witnesses	$> 0.0\%$ of unauthenticated injection payloads are admitted into Ring 1 cohort state
P5. Benign utility vs. noise-free ceiling	At $\varepsilon_{\mathrm{step}} = 0.50$, Concentric DP Meta-DAG retains $\geq 85\%$ benign DAG recovery while trailing noise-free public DAG by ≤ 10 pp	Benchmark S-DAG reasoning aggregation [15] across 16 agents at $\varepsilon_{\mathrm{step}} = 0.50$	Benign DAG consensus recovery drops below 85% or trails noise-free aggregation by > 10 percentage points
P6. Epistemic herding resilience	Diversity-weighted cohort aggregation (3) bounds false-consensus cascades below 10% under correlated errors	Simulate 25% homogeneous model cohorts subject to malignant epistemic herding [16] across 500 ontology nodes	$> 10.0\%$ of Ring 2 ontology edges converge to uncorroborated herding hallucinations

9 Conclusion

Multi-agent provenance ledgers cannot rely on flat public visibility—where internal inter-agent C2 channels leak 68.8% of private attributes—nor on heuristic embedding obfuscation that collapses under sequential cross-query accumulation. By structuring provenance into three concentric rings—Ring 0 on-device W3C PROV-DM chronicles with symbolic indirection, Ring 1 permissioned cohort DAGs gated by Denning/DLM derivation-lattice intersections and SHA-256 Merkle

commitments, and Ring 2 public ontological meta-layers protected by metric differential privacy, Negative-Binomial dummy sensitivity dilution, and a hard privacy odometer cap $\varepsilon_{\max}$—Concentric DP Meta-DAG restricts high-confidence Bayesian attribute reconstruction to 0.67% [0.26, 1.70] across $Q = 160$ and $Q = 640$ queries while retaining 87.17% [84.25, 89.61] benign ontological consensus utility.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Faruk Alpay and Taylan Alpay. AgentSecBench: Measuring prompt injection, privacy leakage, and tool-use integrity in LLM agents. arXiv preprint arXiv:2605.26269, 2026. URL https://arxiv.org/abs/2605.26269.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL https://doi.org/10.1109/CSF.2018.00026.

[3] Prerna Arote, Adrian Saiz, Oriol Saguillo, and Lucianna Kiffer. Cryptographic security is not enough: Privacy gaps in the Renegade decentralized dark pool. arXiv preprint arXiv:2609.27100, 2026. URL https://arxiv.org/abs/2609.27100.

[4] Imen Bajar, Héber Hwang Arcolezi, Antoine Boutet, Carole Frindel, and Mohamed Maouche. Quantifying privacy risks in foundation model representations of brain MRI. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.20494, 2026. URL https://arxiv.org/abs/2610.20494.

[5] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL https://arxiv.org/abs/2605.00914.

[6] Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *USENIX Security Symposium*, 2024. URL https://arxiv.org/abs/2402.06363.

[7] Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. SecAlign: Defending against prompt injection with preference optimization. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2025. URL https://arxiv.org/abs/2410.05451.

[8] Coalition for Content Provenance and Authenticity. The C2PA launches content credentials 2.3 and celebrates 5 years of impact, 2026. URL https://c2pa.org/the-c2pa-launches-c

ontent-credentials-2-3-and-celebrates-5-years-of-impact-across-the-digital-ecosystem/. 9 February 2026.

[9] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[10] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. URL https://arxiv.org/abs/2406.13352.

[11] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[12] Dorothy E. Denning. A lattice model of secure information flow. Communications of the ACM, 19(5):236–243, 1976. URL https://doi.org/10.1145/360051.360056.

[13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014. URL https://doi.org/10.1561/0400000042.

[14] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv preprint arXiv:2602.11510, 2026. URL https://arxiv.org/abs/2602.11510.

[15] Dong et al. S-dag: Subject-based directed acyclic graph for multi-agent heterogeneous reasoning. AAAI Conference on Human Computation and Artificial Intelligence (AAAI 2026), 2026. URL https://arxiv.org/abs/2511.06727.

[16] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[17] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[18] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[19] Joseph A. Goguen and José Meseguer. Security policies and security models. IEEE Symposium on Security and Privacy, pages 11–20, 1982. URL https://doi.org/10.1109/SP.1982.10014.

[20] Enis Golaszewski, Neal Krawetz, Alan T. Sherman, Edward Zieglar, Sai K. Matukumalli, Roberto Yus, Carson L. Kegley, Michael Barthel, William J. Bowman, Bharg Barot, and Kaur Kullman. Verifying provenance of digital media: Why the C2PA specifications fall short. arXiv preprint arXiv:2604.24890, 2026. URL https://arxiv.org/abs/2604.24890.

[21] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pages 31–40, 2007. URL https://doi.org/10.1145/1265530.1265535.

[22] Yu He, Haozhe Zhu, Yiming Li, Shuo Shuo, Hongwei Yao, Zhihao Liu, and Zhan Qin. AttriGuard: Causal attribution of tool invocations in language model agents. arXiv preprint arXiv:2603.10749, 2026. URL https://arxiv.org/abs/2603.10749.

[23] Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting. *arXiv preprint arXiv:2403.14720*, 2024. URL https://arxiv.org/abs/2403.14720.

[24] et al. Jingcheng Wu, Ratan Bahadur Thapa. Sparqlcirc: Probabilistic query evaluation with provenance circuits. arXiv:2611.00128 [cs.DB, cs.AI], 2026. URL https://arxiv.org/abs/2611.00128.

[25] Masahiro Kaneko. Poisoning-based backdoors are not cryptographic. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.28045, 2026. URL https://arxiv.org/abs/2610.28045.

[26] Juhee Kim and Woohyuk Choi. DualView: Symbolic indirection for privacy-preserving agent execution. arXiv preprint arXiv:2607.03821, 2026. URL https://arxiv.org/abs/2607.03821.

[27] Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C Recommendation, 2013. URL https://www.w3.org/TR/prov-dm/.

[28] Andrew C. Myers and Barbara Liskov. A decentralized model for information flow control. Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 129–142, 1997. URL https://doi.org/10.1145/268998.266669.

[29] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[30] Fabio Vitali and Valentina Pasqual. Provenance-enhanced statements in knowledge graphs. arXiv:2606.15246 [cs.AI], 2026. URL https://arxiv.org/abs/2606.15246.

[31] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[32] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. SybilLimit: A near-optimal social network defense against sybil attacks. IEEE Symposium on Security and Privacy (S&P), pages 3–17, 2008. URL https://doi.org/10.1109/SP.2008.13.

[33] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. Findings of the Association for Computational Linguistics (ACL), 2024. URL https://arxiv.org/abs/2403.02691.

[34] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.